Proceedings of the 12th European Conference on Combinatorics, Graph Theory and Applications
EUROCOMB'23
Prague, August 28 - September 1, 2023

The k-XORSAT threshold revisited

(Extended abstract)

Amin Coja-Oghlan* Mihyun Kang† Lena Krieg‡ Maurice Rolvien§

Abstract

We provide a simplified proof of the random k-XORSAT satisfiability threshold theorem. As an extension we also determine the full rank threshold for sparse random matrices over finite fields with precisely k non-zero entries per row. This complements a result from [Ayre, Coja-Oghlan, Gao, Müller: Combinatorica 2020]. The proof combines physics-inspired message passing arguments with a surgical moment computation. MSc: 60B20, 15B52

DOI: https://doi.org/10.5817/CZ.MUNI.EUROCOMB23-041

1 Introduction

A random k-XORSAT instance consists of a conjunction of random XOR clauses with k literals. The goal of the well-known random k-XORSAT problem is to determine the maximum number of XOR-clauses in a random k-XORSAT formula such that the formula remains satisfiable with high probability (w.h.p. for short). This threshold was derived for the random 3-XORSAT problem ($k = 3$) by Dubois and Mandler [14]. They stated that their proof extends to the general case. But this turned out to be far from straightforward.

*TU Dortmund, Faculty of Computer Science, 12 Otto-Hahn-St, Dortmund 44227, Germany. E-mail: amin.coja-oghlan@tu-dortmund.de. Supported by DFG CO 646/3 and DFG CO 646/5

†TU Graz, Institute of Discrete Mathematics, Steyrergasse 30, 8010 Graz, Austria. E-mail: kang@math.tugraz.at. Supported in part by a Friedrich Wilhelm Bessel research award of the Alexander von Humboldt Foundation (AUT 1204138 BES)

‡TU Dortmund, Faculty of Computer Science, 12 Otto-Hahn-St, Dortmund 44227, Germany. E-mail: lena.krieg@tu-dortmund.de. Supported by DFG CO 646/5.

§TU Dortmund, Faculty of Computer Science, 12 Otto-Hahn-St, Dortmund 44227, Germany. E-mail: maurice.rolvien@tu-dortmund.de

Only more than ten years later did Pittel and Sorkin [26] publish the first complete yet complicated proof based on moment computations. Their proof spans well over 30 pages and resorts to computer-assistance. Subsequently, Ayre, Coja-Oghlan, Gao and Müller [4] published a different but still complicated proof based on coupling arguments.

In this work we provide a relatively short proof for the random k-XORSAT satisfiability threshold. Our proof is based on a novel combination of physics-inspired 'quenched' arguments and 'annealed' computations.

We start with a quenched argument. Using a message passing technique called Warning Propagation ('WP') we characterize typical solutions of random k-XORSAT instances. Equipped with this characterization we then carry out a carefully truncated moment calculation ('annealed' computation in physics jargon).

Let $\boldsymbol{F} = \boldsymbol{F}_k(n, m)$ be a random k-XORSAT instance consisting of n Boolean variables and m random XOR-clauses with k literals, where the clauses are drawn independently and uniformly from the set of all possible $2^k\binom{n}{k}$ XOR-clauses of length k on n variables. The k literals of a clause are drawn independently at random. The following theorem, first established in [14] for $k = 3$ and in [26] for $k > 3$, provides the k-XORSAT satisfiablity threshold.

Theorem 1. *For $k \geq 3$ and $d > 0$ let*

$$\Phi_{d,k}(\alpha) = \exp\left(-d\alpha^{k-1}\right) + d\alpha^{k-1} - \frac{d(k-1)}{k}\alpha^k - \frac{d}{k} \qquad \text{and} \tag{1.1}$$

$$d_k = \sup\left\{d > 0 : \max_{\alpha\in[0,1]} \Phi_{d,k}(\alpha) = 1 - d/k\right\}. \tag{1.2}$$

For any $\varepsilon > 0$ w.h.p. the random k-XORSAT formula $\boldsymbol{F}$ is

(i) satisfiable if $m \leq (1-\varepsilon)d_k n/k$, *(ii) unsatisfiable if $m \geq (1+\varepsilon)d_k n/k$.*

A k-XORSAT formula can naturally be translated to a linear system over $\mathbb{F}_2$ and therefore it induces a random matrix over $\mathbb{F}_2$ where each column represents a variable and each row a clause of the formula. Theorem 1 admits a natural generalisation to matrices over finite fields beyond $\mathbb{F}_2$.

Thus, let $q \geq 2$ be a prime power and let $\mathfrak{A} = (\mathfrak{A}_{ij})_{i,j\geq 1}$ be an infinite matrix with non zero entries $\mathfrak{A}_{ij} \in \mathbb{F}_q \setminus \{0\}$. Further, we choose a sequence $(\boldsymbol{e}_i)_{i\geq 1}$ of independent uniformly random subsets of $[n]$ of size $|\boldsymbol{e}_i| = k$. Define the random $m \times n$-matrix $\boldsymbol{A} = \boldsymbol{A}(k, m, n, q, \mathfrak{A})$ over $\mathbb{F}_q$ by letting

$$\boldsymbol{A}_{ij} = \mathfrak{A}_{ij}\mathbb{1}\{j \in \boldsymbol{e}_i\} \qquad (i \in [m],\, j \in [n]).$$

For $q = 2$ we obtain the matrix induced by a k-XORSAT formula.

Theorem 2. *For all $k \geq 3$, all prime powers $q \geq 2$ and all infinite matrices $\mathfrak{A}$ composed of non-zero elements of $\mathbb{F}_q$ the following hold. Let d_k be the threshold from* (1.2)*. Then for any $\varepsilon > 0$,*

(i) if $m \leq (1-\varepsilon)d_k n/k$, then $\boldsymbol{A}$ has full row rank w.h.p.

(ii) if $m \geq (1+\varepsilon)d_k n/k$, then $\boldsymbol{A}$ fails to have full row rank w.h.p.

Theorem 2 complements [4, Theorem 1.1], where only random matrices with identically distributed rows were considered, while in Theorem 2 random matrices may proscribe different non-zero entries for each row. We proceed to outline the proof strategy of Theorem 2.

2 Proof strategy

The main task is to prove the positive statement Theorem 2(i). Assume that for $m < (1-\varepsilon)d_k n/k$ w.h.p. the values of a random kernel vector $\boldsymbol{\sigma} \in \ker \boldsymbol{A}$ are approximately 'balanced' such that each value $s \in \mathbb{F}_q$ appears in $\boldsymbol{\sigma}$ about n/q times. Via a moment calculation we could show that the expected number of such balanced vectors $\sigma \in \ker \boldsymbol{A}$ equals $(1+o(1))q^{n-m}$. Thus $|\ker \boldsymbol{A}| = (1+o(1))q^{n-m}$ w.h.p. and $\boldsymbol{A}$ has full row rank w.h.p. via the second moment method.

Hence, it remains to show that a typical kernel vector $\boldsymbol{\sigma} \in \ker \boldsymbol{A}$ is balanced. However, we are not able to prove directly that a random kernel vector is balanced w.h.p. Instead, we will use a technique called Warning Propagation ('WP') to extract a quantitative picture of the kernels vectors' structure via a 'quenched' argument.

Pinning. We begin with an auxiliary result from [6]. Let A be an $M \times N$ matrix over finite field $\mathbb{F}_q$. We alter the matrix A using a technique called pinning: We add a few rows to the matrix with exactly one non-zero entry at a random position which thus pin the corresponding variables to zero. This randomised pinning operation, devised in this form in [6], mostly removes 'short linear relations' from the matrix and actually works on any arbitrary matrix.

Following [6] we call a set of columns J a *relation of* A if there exists a linear combination of rows with J as the set of non zero entries. Hence J is a relation of A if there exists a vector $y \in \mathbb{F}_q^M$ such that $\mathrm{supp}(y^\top A)$ is a non-empty subset of J. For a k-XORSAT formula these relations can be interpreted as derived XOR-clauses.

Further we call a column or variable *frozen in* A, if the singleton $\{j\}$ is a relation of A. Thus, j is frozen iff every kernel vector is zero on position j. We denote $\mathcal{F}(A)$ as the set of frozen coordinates in A and say that $J \neq \emptyset$ is a *proper relation* of A if $J \setminus \mathcal{F}(A)$ is a relation of A. Finally, we say that A is (δ,ℓ)*-free* if A possesses fewer than $\delta\binom{N}{h}$ proper relations I of size $|I| = h$ for any $2 \leq h \leq \ell$. In other words, a matrix is (δ,ℓ)-free if it contains only few short relations that are not exclusively composed of frozen coordinates.

For an integer $t \geq 0$ let $A[t]$ denote a matrix obtained from A by adding t new rows, each of which contains a single non-zero entry at a random position.

Lemma 1 ([6, Proposition 2.4])**.** *For any $\delta > 0, \ell > 0$ there exists $T_0 = O(\ell^3/\delta^4) > 0$ such that for any $T \geq T_0$ and any matrix A for a random $\boldsymbol{t} \in [T]$ we have $\mathbb{P}[A[\boldsymbol{t}]$ is (δ,ℓ)-free$] > 1-\delta$.*

Thus, with $T = \lceil \log n \rceil$, the matrix $\boldsymbol{A}^\dagger = \boldsymbol{A}[\boldsymbol{t}]$ is (ω^{-1}, ω)-free with $\omega = \lceil \log \log n \rceil$ w.h.p. This allows us to characterize the set of frozen variables in $\boldsymbol{A}^\dagger$ in terms of the Warning Propagation scheme.

Warning Propagation. We introduce WP for a general $M \times N$ matrix A, not just for $\boldsymbol{A}^\dagger$. The matrix A naturally induces a bipartite graph $G(A)$ called the *Tanner graph* with two different kind of vertices, *variable nodes* and *check nodes*. The set of variable nodes and check nodes coincide with the columns and rows of the matrix.

We define the WP scheme following [5]. The goal is to characterize the set of variables frozen in the matrix A in terms of local interactions between variable nodes and their adjacent checks using WP messages. Each edge $v_j a_i$ is endowed with two messages, one sent by the variable node v_j to the factor node a_i and one from the factor node to the variable node. Each message takes a symbolic value $\{\mathtt{u}, \mathtt{f}\}$ to represent 'unfrozen' and 'frozen'.

The *standard messages* $\mathtt{m}_{v_j \to a_i}(A)$ encompass the actual effects of adjacent variables and factors emerging of the matrix. Let $A \setminus \{a_i\}$ be the matrix obtained from A by deleting the row a_i. Similarly $A \setminus \{\partial v_j \setminus \{a_i\}\}$ is the matrix where every other row adjacent to v_j except a_i is removed. The standard message $\mathtt{m}_{v_j \to a_i}(A) = \mathtt{f}$ indicates that the variable v_j is frozen in the matrix $A \setminus \{a_i\}$. Similarly, $\mathtt{m}_{a_i \to v_j}(A) = \mathtt{f}$ expresses that v_j needs to be frozen in order to satisfy the check a_i and thus is frozen in $A \setminus \{\partial v_j \setminus \{a_i\}\}$.

Warning Propagation update provides a heuristic fixed point equation for these messages:

$$\mathtt{m}_{v_j \to a_i} = \begin{cases} \mathtt{f} & \text{if } \exists a_h \in \partial v_j \setminus \{a_i\} : \mathtt{m}_{a_h \to v_j} = \mathtt{f}, \\ \mathtt{u} & \text{otherwise,} \end{cases} \tag{2.1}$$

$$\mathtt{m}_{a_i \to v_j} = \begin{cases} \mathtt{f} & \text{if } \forall v_h \in \partial a_i \setminus \{v_j\} : \mathtt{m}_{v_h \to a_i} = \mathtt{f}, \\ \mathtt{u} & \text{otherwise.} \end{cases} \tag{2.2}$$

The idea is that freezing is caused by local effects only. For instance v_j is expected to be frozen in $A \setminus \{a_i\}$ iff some other check a_h freezes v_j via a standard message.

The fixed point equations (2.1),(2.2) are easily verified for matrices with acyclic Tanner graphs. However, they do not hold for general matrices. Nonetheless, we show that for the random matrix $\boldsymbol{A}^\dagger$ (2.1),(2.2) hold for all but $o(n)$ adjacent pairs a_i, v_j w.h.p. and that the messages correctly identify the set of frozen variables. Furthermore, we prove that in most kernel vectors the values of the unfrozen variables are approximately 'balanced'.

Quenched analysis. Recall that our goal is to show that a random kernel vector $\boldsymbol{\sigma}^\dagger \in \ker \boldsymbol{A}^\dagger$ is approximately balanced w.h.p. Since we know that this holds for the unfrozen variables due to the WP-results, we only need to show that the fraction of frozen variables is $\boldsymbol{\alpha} = o(1)$ w.h.p. For this purpose we will extract detailed quantitative information about combinations of messages belonging to an edge as well as the number of certain labels.

Our next goal is to derive this information in terms of the (as of yet) unknown random variable $\boldsymbol{\alpha}$.

We denote by $\ell = (\ell_{\mathtt{uu}}, \ell_{\mathtt{uf}}, \ell_{\mathtt{fu}}, \ell_{\mathtt{ff}}) \in \mathbb{Z}^4_{\geq 0}$ a specification of message combinations, where $\ell_{\mathtt{uf}}$ equals the number of edges with message combination `u` (incoming) `f` (outgoing), etc. Define Δ_ℓ as the number of variable nodes that receive/send out messages according to ℓ. Analogously, let Γ_ℓ be the number of factor nodes that receive/send according to ℓ.

We are going to estimate $|\Delta_\ell|$ and $|\Gamma_\ell|$ in terms of the fraction $\boldsymbol{\alpha}$ of frozen variables using the hypothesis that the incoming messages at a check node a_i are essentially independent. We can derive predictions $\bar{\Gamma}_\ell(\boldsymbol{\alpha})$ and $\bar{\Delta}_\ell(\boldsymbol{\alpha})$ in terms of the (obvious) Galton Watson tree that mimics the Tanner graph of $\boldsymbol{A}$ and show that these approximations are accurate w.h.p.

Proposition 1. *Let $d > 0, k \geq 3$. Then w.h.p. for all but $o(n)$ adjacent pairs v_j, a_i the fixed point equations* (2.1),(2.2) *hold. Moreover for all ℓ*

$$\mathbb{E}\left||\Delta_\ell| - n\bar{\Delta}_\ell(\boldsymbol{\alpha})\right| + \mathbb{E}\left||\Gamma_\ell| - m\bar{\Gamma}_\ell(\boldsymbol{\alpha})\right| = o(n).$$

Finally, for all but $o(n)$ exceptions variable v_j is frozen iff $\mathfrak{m}_{a_i \to v_j} = \mathtt{f}$ for some $a_i \in \partial v_j$.

The proof of Proposition 1 is based on coupling arguments and does not reveal the likely value of $\boldsymbol{\alpha}$.

Annealed argument. In the next and last step we aim to show that $\boldsymbol{\alpha} = o(1)$ w.h.p. if $d < (1-\varepsilon)d_k$. The present annealed computation differs significantly from the prior works of [14, 26]. These prior works were based on blunt moment computations that generally have the disadvantage that even extremely rare events contribute. These large deviations result in intricate and technically demanding analytical optimisation problems. In contrast, thanks to Proposition 1 we already know the typical shape of kernel vectors and are therefore left with a straightforward and elegant computation.

To elaborate, we proceed in two steps. First, we estimate the expected number of α-WP fixed points with an α-fraction of frozen variables, which turns out to be sub-exponential for any $0 \leq \alpha \leq 1$. In the next step we estimate the number $\boldsymbol{X}_\alpha$ of kernel vectors $\boldsymbol{\sigma}^\dagger \in \ker(\boldsymbol{A}^\dagger)$ that extend a certain α-WP fixed point (frozen variable set to zero, unfrozen variables balanced). Proposition 1 then implies that $|\ker \boldsymbol{A}^\dagger| \sim \boldsymbol{X}_{\boldsymbol{\alpha}}$ w.h.p. Let $\mathfrak{D}$ be the σ-algebra generated by the degree-sequence of the Tanner graph. The following proposition gives a first moment upper bound on $\boldsymbol{X}_\alpha$ for any $0 \leq \alpha \leq 1$ in terms of the function $\Phi_{d,k}$ from (1.1).

Proposition 2. *Let $d > 0, k \geq 3$. W.h.p. for all $\alpha \in [0,1]$ we have*

$$\mathbb{E}[\boldsymbol{X}_\alpha \mid \mathfrak{D}] \leq q^{n\Phi_{d,k}(\alpha)+o(n)}.$$

For $d < d_k$ the function $\Phi_{d,k}$ has its unique maximum at $\alpha = 0$ and $q^{n\Phi_{d,k}(0)} = q^{n-dn/k}$. Thus, we can derive the estimate $\boldsymbol{\alpha} = o(1)$ w.h.p. and finally we can deduce that w.h.p. most kernel vectors $\boldsymbol{\sigma}^\dagger \in \ker(\boldsymbol{A}^\dagger)$ are 'balanced'. This finishes our proof strategy outlined at the beginning.

Acknowledgment Special thanks to Olga Scheftelowitsch for her comments on an earlier version of this work.

References

[1] D. Achlioptas and M. Molloy. The solution space geometry of random linear equations. *Random Struct. Algorithms*, 46:197–231, 2015.

[2] D. Achlioptas, A. Naor, and Y. Peres. Rigorous location of phase transitions in hard optimization problems. *Nature*, 435:759–764, 2005.

[3] M. Aizenman, R. Sims, and S. Starr. An extended variational principle for the sk spin-glass model. *Phys. Rev. B*, 68:214403, 2003.

[4] P. Ayre, A. Coja-Oghlan, P. Gao, and N. Müller. The satisfiability threshold for random linear equations. *Combinatorica*, 40:179–235, 2020.

[5] A. Coja-Oghlan, O. Cooley, M. Kang, J. Lee, and J. Ravelomanana. The sparse parity matrix. *Proc. 33rd SODA*, pages 822–833, 2022.

[6] A. Coja-Oghlan, A. Ergür, P. Gao, S. Hetterich, and M. Rolvien. The rank of sparse random matrices. *Proc. 31st SODA*, pages 579–591, 2020.

[7] A. Coja-Oghlan, P. Gao, M. Hahn-Klimroth, J. Lee, N. Müller, and M. Rolvien. The full rank condition for sparse random matrices. *arXiv:2112.14090*, 2021.

[8] O. Cooley, J. Lee, and J. Ravelomanana. Warning propagation: stability and subcriticality. *arXiv:2111.15577*, 2021.

[9] C. Cooper. The cores of random hypergraphs with a given degree sequence. *Random Struct. Algorithms*, 25:353–375, 2004.

[10] C. Cooper, A. Frieze, and W. Pegden. On the rank of a random binary matrix. *Electron. J. Comb.*, 26, 2019. P4.12.

[11] N. Creignou, H. Daude, and O. Dubois. Approximating the satisfiability threshold for random k-xor-formulas. *arXiv:cs/0106001*, 2001.

[12] M. Dietzfelbinger, A. Goerdt, M. Mitzenmacher, A. Montanari, R. Pagh, and M. Rink. Tight thresholds for cuckoo hashing via xorsat. *Proc. 37th ICALP*, pages 213–225, 2010.

[13] J. Ding, A. Sly, and N. Sun. Proof of the satisfiability conjecture for large k. *Ann. of Math.*, 196:1–388, 2022.

[14] O. Dubois and J. Mandler. The 3-xorsat threshold. *Proc. 43rd FOCS*, pages 769–778, 2002.

[15] D. Fernholz and V. Ramachandran. Cores and connectivity in sparse random graphs. *UTCS Technical Report TR04-13*, 2004.

[16] A. Goerdt and L. Falke. Satisfiability thresholds beyond k-xorsat. *Proc. 7th International Computer Science Symposium in Russia*, pages 148–159, 2012.

[17] M. Ibrahimi, Y. Kanoria, M. Kraning, and A. Montanari. The set of solutions of random xorsat formulae. *Ann. Appl. Probab.*, 25:2743–2808, 2015.

[18] S. Janson and M. Luczak. A simple solution to the k-core problem. *Random Struct. Algorithms*, 30:50–62, 2007.

[19] S. Janson, T. Luczak, and A. Rucinski. Random graphs. *Wiley*, 2000.

[20] J.H. Kim. Poisson cloning model for random graphs. *Proc. International Congress of Mathematicians*, pages 873–897, 2006.

[21] M. Mézard and A. Montanari. Information, physics and computation. *Oxford University Press*, 2009.

[22] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina. Two solutions to diluted p-spin models and xorsat problems. *J. Stat. Phys.*, 111:505–533, 2003.

[23] M. Molloy. Cores in random hypergraphs and boolean formulas. *Random Struct. Algorithms*, 27:124–135, 2005.

[24] A. Montanari. Estimating random variables from random sparse observations. *European Trans. on Telecomm.*, 19(4):385–403, 2008.

[25] D. Panchenko and M. Talagrand. Bounds for diluted mean-field spin glass models. *Probab. Theory Relat. Fields*, 130:319–336, 2004.

[26] B. Pittel and G. Sorkin. The satisfiability threshold for k-xorsat. *Combin. Probab. Comput.*, 25:236–268, 2016.

[27] B. Pittel, J. Spencer, and N. Wormald. Sudden emergence of a giant k-core in a random graph. *J. Combin. Theory Ser. B*, 67:111–151, 1996.

[28] P. Raghavendra and N. Tan. Approximating csps with global cardinality constraints using sdp hierarchies. *Proc. 23rd SODA*, pages 373–387, 2012.

[29] O. Riordan. The k-core and branching processes. *Combin. Probab. Comput.*, 17:111–136, 2008.